Mining Medical Data: Bridging the Knowledge Divide

Sam Schmidt[1], Peter Vuillermin[2], Bernard Jenner[2], Yongli Ren[3], Gang Li[1], and Yi-Ping Phoebe Chen[1]

[1] School of Engineering and Information Technology, Deakin University Australia
[2] The Geelong Hospital, Barwon Health
[3] School of Information Engineering, Zhengzhou University China
sam.schmidt@deakin.edu.au, peterv@barwonhealth.org.au, admin@jenner.com.au, gang.li@deakin.edu.au, phoebe.chen@deakin.edu.au, yonglitom@gmail.com

Abstract. Due to the significant amount of data generated by modern medicine there is a growing reliance on tools such as data mining and knowledge discovery to help make sense and comprehend such data. The success of this process requires collaboration and interaction between such methods and medical professionals. Therefore an important question is: How can we strengthen the relationship between two traditionally separate fields (technology and medicine) in order to work simultaneously towards enhancing knowledge in modern medicine. To address this question, this study examines the application of data mining techniques to a large asthma medical dataset. A discussion introducing various methods for a smooth approach, straying from the 'jack of all trades, master of none' to a modular cooperative approach for a successful outcome is proposed. The results of this study support the use of data mining as a useful tool and highlight the advantages on a global scale of closer relations between the two distinct fields. The exploration of CRISP methodology suggests that a 'one methodology fits all approach' is not appropriate, but rather combines to create a hybrid holistic approach to data mining.

Keywords: medical data mining, data mining, knowledge discovery

1 Introduction

What sets apart medical and generic data? If the data mining techniques applied are the same, then what difficulties slow the process, or prevent it altogether when medical data is in question? Such questions are investigated with evidence based research, such as Lovis et al [1] demonstrating the challenges in medical data mining at 'all levels', taking into account the 'technical', 'semantical', 'legal' and 'ethical' issues arising from such work.

Modern medicine today can generate raw text computerised output, as well as two or three dimensional images, all of which can have various data mining algorithms applied for results. Apart from business, where data mining has some history it can also be used to obtain key results in the field of medicine, within

areas such as bioinformatics and biomechanics for predictions or diagnostic purposes. It consequently allows for enormous benefits to be gained if the data can be presented in ways which allow for the generation of previously unseen linkages, or new clusterings of information.

Fayyad et al [2]defines knowledge discovery as 'the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data'. Knowledge discovery is therefore holding increasing relevance to medical areas, and this is frequently being acknowledged within the research literature. For example, Kudyba [3]states that 'strong disease management programs depend on data mining techniques', and he also refers specifically to asthma. Given the vast amounts of medical data available today, data mining and artificial intelligence techniques appear to offer the best method of knowledge discovery. Data mining may allow for the identification of natural groups of patients given inputs such as symptoms, and further classify or predict derivatives from the given data. Within the setting of this study, the application of various supervised and unsupervised data mining techniques constitute the knowledge discovery process.

In the past, examining large sets of biomedical data has been a tedious and sometimes impossible task. Now, computational techniques offer a potential solution to such an obstacle, allowing such data sets to be examined in ways that have previously not been available. Therefore, the presence of computer science in the medical realm is increasing [4]. The presence of computer science in the medical realm is increasing. The ramification of the merging of these fields is far reaching, particularly in the areas of epidemiology, prevention and public health overall. As a consequence, it is important that the differences among these fields are considered, along with the possible implications of such differences in the research setting. For example, people working in these fields in most cases have different training, knowledge, approaches and interests, all of which differences may foster scientific miscommunication.

Data Mining has been an area of emerging technology over the past decade, picking up speed and growing rapidly as various industries realise the potential of such practice. Since the inception of this relatively new area, work flows and procedures have emerged; perhaps the key methodology is CRISP-DM, the *Cross Industry Standard Platform for Data Mining* as followed by longstanding statistical software applications such as SPSS (Clementine).

To demonstrate other such issues which arise when working across professions, many barriers are immediately introduced when heterogenous medical data is involved. Such barriers include the ethical, legal and social issues relating to the use of the data. The 'uniqueness' of medical data mining, as described by Moore and Krzysztof [5] may be categorically defined into four groups: a) *Heterogeneity of medical data*, b) *Ethical, legal, and social issues*, c) *Statistical philosophy* and d) *Special status of medicine*. These groups present problems and barriers, perhaps due to traditional training methods that thwart the bridging of the knowledge gap. These four groups demonstrate obstacles that need to be overcome for a successful outcome, a process which may often be time consuming.

As has been the case with the data set under examination in the current study, a traditional statistical philosophy has been applied, but no application of data mining or unassisted knowledge discovery techniques have been performed. From this, one might surmise that the traditional statistical approach is used in the majority of medical data analysis. One major advantage of data mining over a traditional statistical approach is its ability to deal directly with heterogenous data fields, which are usually contained in medical data sets [6]. Therefore, it remains unclear as to whether the results found from the statistical approach are in fact suboptimal in comparison with undirected or unsupervised knowledge discovery techniques. Such conventional analysis remains extremely useful and important; however as demonstrated by Bayat et al [7] it does not allow for such scope as that of data mining, which takes a more 'globalised' view of the data and the particular associations pertaining within. This increases 'knowledge discovery' and may uncover hidden information among data.

Interestingly, Setoguchi et al [8] recently demonstrate the merit of data mining in an evaluation paper of techniques, showing that data mining resulted in the least bias estimates and more accuracy than traditional approaches. To finish the introduction with a real life example also demonstrating the utility of data mining Pearl et al [9] illustrated the use of neural networks to predict patient treatment on the scene by paramedics. Paramedics need to make a life decision, instantly choosing between treatment or 'scoop and run'. Such important decisions can now be assisted with machine learning data mining techniques.

The following sections outline various existing methodologies, and examine in further detail particular techniques for the discovery of knowledge. Currently, in Australia there is a distinct lack of research that has specifically explored the application of such techniques to data sets concerning the condition of asthma. This is a process undertaken in the current study.

2 Methodology

This study initially followed the Cross Industry Standard Practice for Data Mining (CRISP-DM) methodology [10]. As one of the initial workflows contained in the methodology, data understanding requires careful consideration and attention. In order for this understanding to flow smoothly, communication between the fields of computer science and medicine is imperative. In this study, the knowledge gap was lessened and understanding gained, by close collaboration and consultation with the domain experts. This involved inspection of the data set by the professionals from both fields. This enabled clarification and understanding of the data set to occur from the distinct perspective of each field (i.e. computer science and medicine). As part of this process, data was screened and transformed. A new field 'hospitalisation' was created, which combined two existing categories; 'visits to a hospital due to asthma (hospadm)' and 'visits to a emergency department due to asthma (emerg)'. This enabled all three variables (hospitalisation, hospadm and emerg) to be investigated as an additional method of knowledge discovery.

In doing so, we also pose the question: can the CRISP-DM methodology be templated atop any dataset and achieve appropriate outcomes? In our case the methodological 'blueprint' soon deviated off path morphing to a hybrid of a holisitic approach. The main reason for this was due to the limitation of scope in particular areas of the aforementioned methodology, in particular, items such as graphing data structure for comprehensibility between professions.

The data set utilised in this study consisted of random sampling conducted at primary schools in the Barwon region of Victoria, Australia with 16,957 school aged children being surveyed via their parent or guardian [11]. Survey questions related specifically to the symptoms and prevalence of wheezing and asthma experienced by children. To demonstrate the knowledge divide, this dataset has been analysed in depth using traditional statistical techniques, with the major findings from such analyses being a link between medical facility usage and the age of children, and the socio economic status not being associated with the availablity or usage of medical services [11]. In comparison, in the present study we demonstrate knowledge discovery through data mining without any initial hypothesis to prove, or 'traditional' questions to ask. Data mining draws more from the field of machine learning in comparison to statistics, and can therefore be advantageous [6].

Despite the existence of a large body of research in the field of data mining, significant variation exists in relation to the varying processes and naming conventions. These differences have important implications for research practices and make it difficult for researchers to compare and interpret studies. In order to avoid such confusion the current study therefore explicitly defines the following phases: a) data analysis as the understanding, description, exploration and structural visualisation of the initial data; b) data cleaning as the identification and preparation of data for manipulation, and; c) data transformation as the implementation of identified fields within the data cleaning identification process.

3 Case Study Experiment

3.1 Statistical Analysis

Prior to conducting the major analysis, the prevalence estimates of asthma in the current data set were compared to exiting research. Of the 16,957 children surveyed in the asthma data set, 27% were reported as having asthma, and 334 had been hospitalised due to their asthma. This proportion was higher than the 14-16% of children cited as having the disease by Government reports [12]. Of these 27% of children, doctor visitations constituted the majority of the corpus, with hospitalisation (including emergency admissions) accounting for 12% of cases. At 12%, this figure appears to be somewhat low by International standards, i.e. when compared to research conducted in Canada which found 21.2% of all confirmed asthma instances required hospitalisation [13].

Consistent with previous research that also has examined similar age groups [14], the current study found emergency department admissions to be higher than that

of hospitilistaions. While research by Adams et al, 2001 found that emergency department visits accounted for three quarters of hospiatilsation in childhood asthma, in the current study, the emergency department visits accounted for only double the hospital visits.

The survey used in collecting the data for the current study followed the ISAAC (international Study of Asthma and Allergies in Childhood) protocol. In the current study, parents were required to answer the question 'Has your child ever had asthma?'. Previous research on the topic has asked parents to answer questions such as 'Do you have asthma as diagnosed by a health professional?' [15]. While either question allows for discrepant reporting by the informants, it is a difficult methodology question to address. In the current study the accuracy of the information is therefore dependent on the parent's interpretation of 'asthma', and their ability to accurately identify the condition in their child. In addition, there is some research evidence to suggest that variations in the diagnosis of asthma also occurs by medical practitioners. For example, doctors who class bacterial respiratory infections (such as bronchitis) as asthma, which in the strict definition of asthma is incorrect [16]. The restrospective nature of the survey methodology, for example 'Has your child ever had asthma', may potentially affect the accuracy of the information that parents report, and therefore should be noted. Therefore, the data set initially may be slightly skewed. However, the instances of hospitalisation assists in affirmation of a child having 'asthma' per se, as it is not reliant on the parental perceptions.

3.2 Data Mining Techniques

Of the various pieces of knowledge discovered, some examples of the benefits can be shown by: Research Instrument Design - More implicitly, knowledge discovery through the use of Kohonen's SOM shows a high correlation between two variables: nightcough, and sleepdisturbance. Such a result, achievable also through statistical methods, is highly comprehensible and identifiable through SOM, providing instant knowledge discovery rather than 'trial and error' techniques to find such patterns.

Various techniques were used for the prediction of hospitalisation amongst parent reported asthma in children. The following section outlines the techniques used for segmentation, specifically detailing the use of self organising maps for both comprehensibility and exploratory investigations.

3.3 Segmentation

Kohonen's Self Organising Maps (SOM's) were used for data segmentation. Based upon a two dimensional neural network model, SOM's segregate and cluster the data according to similarities within attributes. A major advantage of using SOM's is the high comprehensability available from inspection of the output. The SOM's were accomplished through Kohonen's algorithm presented below:

$$\Delta \boldsymbol{w_i} = c_i(\boldsymbol{x} - \boldsymbol{w_i}^{old}), \ for \ i \in N_r \tag{1}$$

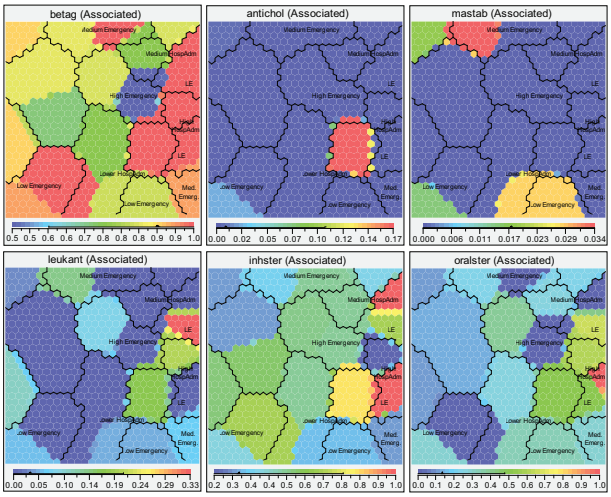

Fig. 1. Six main asthma medication groups associated against hospitalisation

To begin with, medical visitations were examined and clustered via the use of SOM. The attributes that were under examination were: 'hospital admissions (hospadm)' and 'emergency admissions (emerg)', which were weighted with priority. Subsequently, the self-organising map works by placing these attributes together as closely as possible.

Figure 1 shows differences between the six main asthma medication groups: (*Beta Agonists, Anti Cholinergics, Mast Cell Stabilisers, Luekotrienes, Inhaled Steroids* and *Oral Steroids*) and hospitalisation rates. The figure presents a visual representation of the data set with these attributes.

Inspection of Figure 1 reveals that the trends for hospitalisation varied according to the six medication groups. Specifically, trends for children using the various medications for asthma management are described as follows: a) Beta agonists: low emergency/high hospitalisation, b) anti cholinergics: trend for low hospitalisation, c) mast cell stabilisers: high emergency/low hospitalisation, d) leukotrienes: emergency with no repeats, e) inhaled steroids: high hospitalisation/low emergency, and f) oral steroids: high hospitalisation/low emergency.

Further implicit knowledge was gained for future research instrument design, such as two symptomatic variables from the asthma data set, 'sleep disturbance' and 'night cough'. It was found that all instances of night cough are accompanied by sleep disturbance. Whilst a statistical method can also pick up such correlations when directed, the benefits of SOM is the undirected knowledge discovery and comprehensibility, allowing for ease of recognition of such knowledge.

A logical assumption may then be drawn, considering that if a child is coughing at night, their sleep may be disturbed, and vice verse. Therefore, such implicit knowledge discovery can prove to be highly beneficial in the design of future research instruments, and also aid in the diagnosis of asthma. This shows the accuracy of such questions may be skewed and ideally the two questions may perhaps be merged into a single question for a higher success rate and for better evidence based practice.

3.4 Association

The apriori algorithm was used to discover the affiliated attributes with the incidences of hospitalisation. This algorithm was chosen primarily due to the speed of application. While it was originally developed for 'basket data' (i.e. product sales) by IBM, it has since proven to be a useful and robust algorithm across a range of areas [17]. To start with, independent test on both hospital admission and emergency admissions were conducted. While confidence levels were low and therefore not reported, it is noteworthy that in both accounts the antecedents found were the same. These were: wap (written asthma plan), drforwhz (dr for wheezing), medforwhz (medication for wheezing), everast (ever had asthma), and evwhz (ever wheezed).

Examination of the resulting output suggested that everast and evwhz were both obvious indicators. In comparison, the attributes wap, drforwhz and medforwhz were not so obvious indicators. To further discover the associations, a new field 'hospitalisation' was derived by combining both the 'emergency department visits' and 'hospital visits', and the association was again completed. As anticipated, while this still produced low confidence and support levels, they were marginally higher. The weakest indicators had high confidence ratings which therefore assisted in the elimination process of predicting hospitalisation amongst childhood asthma cases. The overall apriori associations are presented in Table 1.

Table 1. Apriori association: contrasting affirmative hospitalisation

Consequent	Antecedent	Support %	Confidence %
hospitalisation = 0	wheezing1to3 = 1 and drforwhz4 = 0 and drforast4 = 0 and evwhz = 1	11.641	94.174
hospitalisation = 0	wheezing1to3 = 1 and drforwhz4 = 0 and evwhz12m = 1	11.824	93.965
hospitalisation = 0	wheezing1to3 = 1 and drforwhz4 = 0 and drforast4 = 0 and everastr = 1 and evwhzr = 1	8.48	93.88
hospitalisation = 0	wheezing1to3 = 1 and drforwhz4 = 0 and medforwhzr = 1 and drforast4 = 0 and everastr = 1	7.643	93.519

As indicated in Table 1, the high confidence rate of children not being hospitalised suggests that the child has the lowest wheezing marker, and has visited the doctor for wheezing or asthma less than 3 times. Likewise, when associating both positive and negative incidences of hospitalisation, commonalities existed. This was particulary evident with the attribute *wap*, perhaps suggesting that this attribute may be a good indicator of hospitalisation.

3.5 Classification

To determine the likelihood of a child being hospitalised due to asthma the current study utilised classification methods. The results from these methods were converted to produce highly comprehensible and structured tree graphs for interpretation. As a result, using the knowledge gathered from the application of the aforementioned methods, a classification algorithm was performed and a tree diagram developed using the data. This enabled ease of inspection by providing a visual diagram of the results. These results are presented in Figure 2.

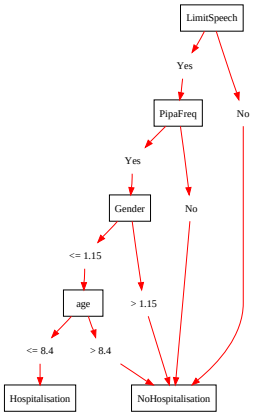

Fig. 2. The QUEST Decision Tree to predict hospitalisation in childhood asthma.

Classification was executed using the binary split decision tree algorithms: QUEST (Quick, Unbiased and Efficient Statistical Tree), CHAID (Chi-Squared Automatic Interaction Detector), C&R and C5.0. The output models generated from the algorithms were examined for both accuracy and comprehensibility. Figure 2 indicates that when using an 80/20 training/test data split, the QUEST algorithm produced a high accuracy, with a comprehendible amount of branches. To enhance the simplicity of the model to enable easy of interpretation for both professions the model was converted to a graphical format. This was done using DOT language, as opposed to isolating it to 'IF A — B THEN C' type structuring.

4 Conclusion

Can computer scientists learn from statisticians who have managed to bridge the knowledge divide, especially within methodological processes such as data analysis and understanding? Whilst a fine line is drawn between statistics and data mining, with both fields borrowing techniques and algorithms from the other, it is important to note there is in fact a clear difference, both fields being

unique but serving differently the same goal. After an initial statistical approach, our results indicate that the data set which had previously been 'cleaned' was able to be further cleaned using various methods: mainly local rule derivation from the initial research instrument which resulted in a number of records being culled for inaccuracy. Obstacles such as these must be streamlined into a fast process will allow for much greater knowledge discovery, ultimately beneficial to the health care network.

This research has provided preliminary evidence to suggest that knowledge discovery techniques such as data mining may indeed uncover different information than traditional statistical methods. The results suggest that such methods warrant further research attention to ascertain the potential benefits and reliability of such methods. As data is generated by modern medicine it is vital that methods of discovering knowledge within such data pools be examined. While traditionally, the fields of medicine and technology have been relatively separate, the increasing overlap has now never been greater. The current research has highlighted a number of reasons for why it is important that the two fields co-operate and open communication channels. Strengthening the ties between these two fields holds enormous potential for significant contributions to be made in both areas.

This study has illuminated the barriers to overcome for a successful outcome. Moreover, it demonstrates the involvement required by many professionals in different industries; without domain expert involvement, accuracy cannot be credibly gauged. The knowledge discovery process has clearly demonstrated important implications for researchers in medical fields. In particular, Kohonen's Self Organising Maps (SOM) identified a range of correlations. While the results of this raised the possibility that seemingly implicit attributes were perhaps suboptimal in the diagnosis of asthma, future research is needed to examine such findings. Many such findings are not possible with traditional statistical methods. For example, the discovery that sleep disturbance and coughing during the night were highly correlated, and also implicitly and logically related to one another. Such results are not possible with traditional methods, unless specifically directed to do so. With a data mining approach, derived mainly from artificial intelligence and machine learning [6], uncovering results shows the need for further discussion with appropriate professionals, which also strengthens linkages between fields. Visualisation may also be a key factor in closing the knowledge divide between medical and computing professionals.

In this article we have proposed that the merging of two traditionally separate fields warrants the use of a system that will allow professionals from both fields to collaborate and work together. Recently Krieger et al [18] outlined the future trends in data mining, and identifies the issue of increasing usability by making data mining methods more user friendly. Such publications strengthen the ties between the computer science and medical professions. From this research it becomes evident, that knowledge which may be discovered, might be vital to future health on a global scale.

References

1. Lovis C, Colaert D, S.V.: (Debugit for patient safety - improving the treatment with antibiotics through multimedia data mining of heterogeneous clinical data)
2. Fayyad U, Piatetsky-Shapiro G, S.P.: Knowledge discovery and data mining: Towards a unifying framework. In: Proceedings of the 2nd ACMinternational conference on knowledge discovery and data mining (KDD), Portland, OR, USA (1996) 82–88
3. Kudyba, S.: Managing Data Mining: Advice from Experts. (2004)
4. Daniel Berleant, t.d.: Medical Informatics: Knowledge Management and Data Mining in Biomedicine. Springer Science and Business Media (2005)
5. Krzysztof J. Cios, G.W.M.: Uniqueness of medical data mining. Artifical Intelligence in Medicine **26** (2002) 1–24
6. Cios, K.J.: Medical data mining and knowledge discovery. IEEE Engineering in Medicine and Biology **7** (2000) 15–16
7. Bayat S, Cuggia M, K.M.B.S.L.B.P.F.L.: Modelling access to renal transplantation waiting list in a french healthcare network using a bayesian method. (Studies in Health Technology and Informatics)
8. Setoguchi S, Schneeweiss S, B.M.G.R.C.E.: (Evaluating uses of data mining techniques in propensity score estimation: a simulation study.)
9. Pearl A, Bar-Or R, B.O.D.: (An artificial neural network derived trauma outcome prediction score as an aid to triage for non-clinicians)
10. Shearer, C.: The crisp-dm model: The new blueprint for data mining. Journal of Data Warehousing **5** (2000)
11. Vuillermin, Peter J. South, M.C.J.B.B.M.I.B.S.L.R.C.F.: Asthma among school children in the barwon region of victoria. Medical Journal of Australia **187** (2007) 221–224
12. Government, A.: Chronic respiratory diseases in Australia Their prevalence, consequences and prevention. Australian Government (2007)
13. Suissa, S; Ernst, P.K.A.: Regular use of inhaled corticosteroids and the long term prevention of hospitalisation for asthma. Thorax **57** (2002) 880–884
14. Robert J. Adams, Anne Fuhlbrigge, J.A.F.P.L.J.M.L.K.B.W., Weiss, S.T.: Impact of inhaled antiinflammatory therapy on hospitalization and emergency department visits for children with asthma. PEDIATRICS **107** (2001) 706–711
15. Yue Chen, Robert Dales, D.K.: Asthma and the risk of hospitalization in canada: The role of socioeconomic and demographic factors. American College of Chest Physicians **119** (2001) 708–713
16. Deirdre Donnelly, Anita Critchlow, M.L.E.: Outcomes in children treated for persistent bacterial bronchitis. Thorax **62** (2007) 80–84
17. Agrawal Rakesh, R.S.: Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB Conference, Santiago, Chile (1994) 487–499
18. Hans-Peter Kriegel, Karsten M. Borgwardt, P.K.A.P.M.S.A.Z.: Future trends in data mining. Data Mining and Knowledge Discovery **15** (2007) 85–97